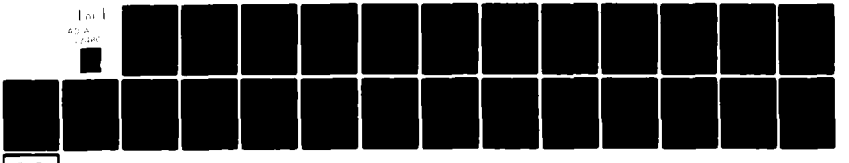
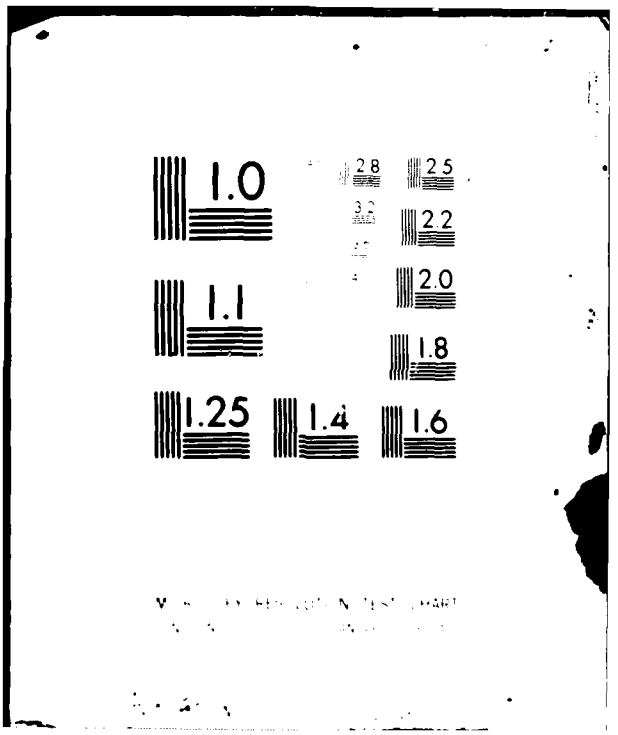
AD-A117 460 TEXAS A AND M UNIV. COLLEGE STATION INST OF STATISTICS F/G 12/1
QUANTILES, PARAMETRIC-SELECT DENSITY ESTIMATIONS, AND BI-INFORM--ETC(U)
JUN 82 E PARZEN DAAG29-80-C-0070
UNCLASSIFIED TR-8-6 ARO-16992.12-MA NL



END
DATE FILMED
08-82
DTIC


1.0
2.8
2.5
3.2
2.2
2.0
1.1
1.8
1.25
1.4
1.6

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

QUANTILES, PARAMETRIC-SELECT DENSITY ESTIMATIONS, AND BI-INFORMATION PARAMETER ESTIMATORS

Emanuel Parzen
Institute of Statistics, Texas A&M University

Technical Report No. B-6

June, 1982

Texas A&M Research Foundation
Project No. 4226

"Robust Statistical Data Analysis and Modeling"

Sponsored by the U.S. Army Research Office
Grant DAAG29-80-C-0070

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Technical Report B-6	2. GOVT ACCESSION NO. AD-A117460
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Quantiles, Parametric-select Density Estimation, and Bi-information Parameter Estimators	5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Emanuel Parzen	8. CONTRACT OR GRANT NUMBER(s): DAAG29-80-C-0070
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas A&M University, Institute of Statistics, College Station, TX 77843	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Army Research Office	12. REPORT DATE: June 1982
	13. NUMBER OF PAGES: 25
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Statistical data science, functional inference, information divergence of index α, bi-information, comparison distribution functions, density estimation.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper outlines a quantile-based approach to functional inference problems in which the parameters to be estimated are density functions. Exponential models and autoregressive models are approximating densities which can be justified as maximum entropy fo respectively the entropy of a probability density and the entropy of a quantile density. It is proposed that bi-information estimation of a density function can be developed by analogy to the problem of identification of regression models.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

CONTENTS

1. Statistical Science, data analysis, and Buffalo snowfall

Statisticians complain about the failure of universities to adequately educate students on how to analyze statistical data. At the same time some statisticians state that data analysis is an art, and thus cannot be taught. When these statisticians speak of statistical science it is difficult to imagine to what they are alluding since they seem to sneeringly reject all attempts to reason, and reach consensus, about the evaluation of methods to be used as part of the process of statistical data analysis.

I would like to propose a data set which I believe provides a useful test case for various approaches to data analysis, namely the annual time series of snowfall in Buffalo, N.Y. The segment of that series which I will discuss is 1910-1972, although it has many interesting features when extended to 1981. The data analysis question to be considered is: What probability distributions can be used to describe Buffalo snowfall. An ever-present hypothesis to be considered is whether Buffalo snowfall is normal.

2. Functions that describe probability distributions

The probability law of a continuous random variable X can be described by one or more of the following functions:

(1) Distribution Function $F(x) = Pr\ [X \leq x]$

(2) Probability Density Function $f(x) = F'(x)$

(3) Quantile Function $Q(u) = F^{-1}(u)$

$= \inf \{x: F(x) \geq u\}$

$= \inf \{x: F(x) = u\}$ if F is continuous

$= x$ such that $F(x) = u$ if F increasing at x

(4) Quantile-Density Function $q(u) = Q'(u)$

(5) Density-Quantile Function $fQ(u) = f(Q(u))$

Theorem: For F continuous

$$FQ(u) = u \quad , \quad fQ(u)\, q(u) = 1$$

3. Raw functions that describe samples

Data $X_1, \ldots, X_n$ is called a random sample of X when $X_1, \ldots, X_n$ are independent random variables identically distributed as X. An important role in the analysis of a sample is played by the order statistics $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$

(1) Sample Distribution $\tilde{F}(x)$ = fraction $X_1, \ldots, X_n \leq x$

$$= \frac{j}{n} \ , \ X_{(j)} \leq x < X_{(j+1)}$$

(2) Sample Probability Density, or Histogram, estimates f(x) by a numerical derivative

$$\tilde{f}(x) = \frac{\tilde{F}(x+h) - \tilde{F}(x-h)}{2h}$$

(3) Sample Quantile $\tilde{Q}(u) = \tilde{F}^{-1}(u)$

$$= X_{(j)}, \ \frac{j-1}{n} < u \leq \frac{j}{n}$$

A universal display of any data set is provided by the quantile box plot introduced in Parzen (1979).

(4) Sample Quantile-Density is a numerical derivative

$$\tilde{q}(u) = \frac{\tilde{Q}(u+h) - \tilde{Q}(u-h)}{2h}$$

(5) Sample Density-Quantile = $\tilde{\tilde{fQ}}(u) = 1/\tilde{q}(u)$.

An important formula is

$$\tilde{f}(X_{(j)}) = \tilde{\tilde{fQ}}(\tfrac{j}{n+1}) = 2\{(n+1)(X_{(j+1)} - X_{(j-1)})\}^{-1}$$

4. Smooth functions that describe samples and estimate probability distributions

The functions F, f, Q, q, fQ that represent the true probability distribution of a random variable X are estimated by smooth functions $\hat{F}$, $\hat{f}$, $\hat{Q}$, $\hat{q}$, $\hat{f}\hat{Q}$ which are derived from the raw descriptive functions $\tilde{F}$, $\tilde{f}$, $\tilde{Q}$, $\tilde{q}$, $\tilde{\tilde{fQ}}$. One distinguishes between parametric and non-parametric methods of estimating smooth functions.

A parametric estimation method : (1) assumes a family F_θ, f_θ, Q_θ, q_θ, $f_\theta Q_\theta$ of functions, called parametric models, which are indexed by a parameter $\theta = (\theta_1, \ldots, \theta_k)$; (2) forms estimators $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ of θ; (3) forms smooth functions by

$$\hat{F}(x) = F_{\hat{\theta}}(x), \quad \hat{f}(x) = f_{\hat{\theta}}(x),$$

$$\hat{Q}(u) = Q_{\hat{\theta}}(u), \quad \hat{q}(u) = q_{\hat{\theta}}(u),$$

$$\hat{f}\hat{Q}(u) = f_{\hat{\theta}}Q_{\hat{\theta}}(u).$$

A non-parametric estimation method forms estimators which are not based on parametric models. Important examples of non-parametric estimators of a probability density f(x) and a

quantile-density q(u) are respectively

$$\hat{f}(x) = \frac{1}{\delta} \int_{-\infty}^{\infty} K(\frac{x-y}{\delta}) \quad d\tilde{F}(x)$$

$$\hat{q}(u) = \frac{1}{\delta} \int_{0}^{1} K(\frac{u-t}{\delta}) \quad d\tilde{Q}(u)$$

for suitable kernels $K(\cdot)$ and bandwidth δ.

5. Parameter estimation and information divergence

When a parametric model f_θ is assumed, parameter estimators $\hat{\theta}$ are often determined by minimizing a "distance" between $\tilde{f}(x)$ and $f_\theta(x)$. A "distance" between two probability densities $f(x)$ and $g(x)$ is denoted $I(f;g)$ and is called an information divergence between $f(x)$ and $g(x)$. It is usually not symmetric in f and g. It does not satisfy the triangle inequality for a metric. But it does satisfy $I(f;g) \geq 0$ and $I(f;g) = 0$ if and only if $f = g$.

The most famous, and most important, definition of information divergence is

$$I_1(f;g) = \int_{-\infty}^{\infty} - \log\{\frac{g(x)}{f(x)}\} \, f(x) \quad dx$$

called the information divergence of order 1, or Kullback-Liebler information divergence. Information divergence of order α is defined for $\alpha>0$ (but $\alpha \neq 1$) by

$$I_\alpha(f;g) = \frac{-1}{1-\alpha} \log \int_{-\infty}^{\infty} \{\frac{g(x)}{f(x)}\}^{1-\alpha} f(x) \, dx.$$

The most important values of α are $0.5 \leq \alpha \leq 2$.

Bi-information divergence is defined by

$$II(f;g) = \int_{-\infty}^{\infty} |\log \{\frac{g(x)}{f(x)}\}|^2 \, f(x) \, dx;$$

it may be regarded as related to $I_2(g;f)$.

Information divergence of order 1 has an important decomposition:

$$I_1(f;g) = H(f;g) - H(f)$$

defining

$$H(f;g) = \int_{-\infty}^{\infty} \{-\log g(x)\}\, f(x)\, dx,$$

$$H(f) = H(f;f) = \int_{-\infty}^{\infty} \{-\log f(x)\}\, f(x)\, dx.$$

We call $H(f;g)$ the cross-entropy of f and g, and call $H(f)$ the entropy of f.

Maximum likelihood parameter estimation can be shown to be equivalent to minimum cross-entropy estimation. The likelihood function of a parametric model f_θ is defined by

$$L(f_\theta) = \log f_\theta(X_1,\ldots,X_n)$$

$$= \sum_{t=1}^{n} \log f_\theta(X_t)$$

One may verify that

$$L(f_\theta) = n \int_{-\infty}^{\infty} \log f_\theta(x)\, d\tilde{F}(x)$$

$$= -n\, H(\tilde{f};\, f_\theta).$$

The maximum likelihood parameter estimator $\hat{\theta}$, defined by

$$L(f_{\hat{\theta}}) = \max_\theta L(f_\theta) \quad ,$$

clearly satisfies

$$H(\tilde{f};f_{\hat{\theta}}) = \min_\theta H(\tilde{f};f_\theta).$$

It also satisfies

$$I_1(\tilde{f};f_{\hat{\theta}}) = \min_\theta I_1(\tilde{f};f_\theta).$$

In general parameter estimators $\hat{\theta}$ are found by minimizing $I_\alpha(\tilde{f};f_\theta)$ or $I_\alpha(f_\theta;\tilde{f})$. Chi-squared estimators minimize $I_2(f_\theta;\tilde{f})$ while modified chi-squared estimators minimize $I_2(\tilde{f};f_\theta)$.

To compute $I_1(\tilde{f};f_\theta)$ one needs to compute $H(\tilde{f})$. A useful formula for accomplishing this is

$$H(f) = \int_{-\infty}^{\infty}\{-\log f(x)\}\, dF(x)$$

$$= \int_0^1 \{-\log fQ(u)\}\, du$$

$$= \int_0^1 \log q(u)\, du.$$

The value of $I_1(\tilde{f};f_{\hat{\theta}})$ can be used to test the goodness of fit of the parametric model f_θ.

6. Information and bi-information parameter estimation, and comparison distribution functions

Given a sample with sample probability density function $\tilde{f}$ and parametric model f_θ, one can form diverse parameter estimators, denoted $\hat{\theta}$ and $\overset{\vee}{\theta}$, corresponding to two choices of information divergence which we take to be: (1) $I_1(\tilde{f};f_\theta)$, and (2) $I_2(f_\theta;\tilde{f})$ or $II(\tilde{f};f_\theta)$. We call $\hat{\theta}$ and $\overset{\vee}{\theta}$ diverse parameter estimators. For greater precision we call $\hat{\theta}$ the (order 1) information estimator, and $\overset{\vee}{\theta}$ the bi-information estimator.

When the parametric model f_θ is exact, the diverse parameter estimators have equivalent statistical properties; they are both asymptotically efficient estimators, and are not significantly different from each other.

When the values of $\hat{\theta}$ and $\overset{\vee}{\theta}$ computed from a sample are significantly different one should suspect that the parametric model f_θ does not fit the data. The Shapiro-Wilk statistics

for testing normality and exponentiality can be regarded as comparing diverse estimators which minimize information of order 1 and 2 respectively.

One can interpret $\hat{\theta}$ and $\overset{\vee}{\theta}$ as parameter values of "best approximating" models.

One wishes to evaluate $F_{\hat{\theta}}(x)$ and $F_{\overset{\vee}{\theta}}(x)$ as smooth estimators of $F(x)$. For any parameter value θ, define

$$\tilde{D}_{\theta}(u) = F_{\theta}(\tilde{Q}(u))$$

which is the sample quantile function of the transformed random variables

$$U_1 = F_{\theta}(X_1), \ldots, U_n = F_{\theta}(X_n).$$

The true parameter value θ has the property that $U_1, \ldots, U_n$ are distributed with a uniform [0,1] distribution. Then parameter estimators $\hat{\theta}$ and $\overset{\vee}{\theta}$ are compared by the character of the closeness to the identity function $D(u) = u$ of $\tilde{D}_{\hat{\theta}}(u)$ and $\tilde{D}_{\overset{\vee}{\theta}}(u)$.

We call $\tilde{D}_{\theta}(u)$ a comparison distribution function. Its derivative

$$\tilde{d}_{\theta}(u) = \{\tilde{D}_{\theta}(u)\}'$$

plays a basic role and is called a comparison density; formulas for the comparison density are

$$\tilde{d}_{\theta}(u) = f_{\theta}(\tilde{Q}(u)) \quad \tilde{q}(u)$$

$$= \frac{f_{\theta}(\tilde{Q}(u))}{\tilde{f}\ \tilde{Q}(u)}$$

An alternative comparison density introduced in Parzen (1979), is

$$\tilde{d}(u) = f_0Q_0(u)\ \tilde{q}(u) \div \tilde{\sigma}_o,$$

$$\tilde{\sigma}_o = \int_o^1 f_oQ_o(u)\ \tilde{q}(u)\ du,$$

$$\tilde{D}(u) = \int_o^u \tilde{d}(t)\ dt$$

where $f_oQ_o(u)$ is a specified density-quantile function.

Parameter estimators can be justified as minimizing information divergence

$$I_1(\tilde{d}_\theta) = \int_o^1 -\log \tilde{d}_\theta(u)\ du = I_1(\tilde{f};f_\theta)$$

$$II(\tilde{d}_\theta) = \int_o^1 |\log \tilde{d}_\theta(u)|^2\ du = II(\tilde{f};f_\theta)$$

$$I_\alpha(\tilde{d}_\theta) = \frac{-1}{1-\alpha} \log \int_o^1 \{\tilde{d}_\theta(u)\}^{1-\alpha} du$$

$$\int_o^1 |\tilde{d}_\theta(u) - 1|^2\ du = \int_o^1 |\tilde{d}_\theta(u)|^2\ du - 1 \quad .$$

These measure the closeness to 1 of $\tilde{d}_\theta(u)$, or the closeness to $D(u) = u$ of $\tilde{D}_\theta(u)$. However the final decision about parameter estimators should be based on visual inspection of the graph of $\tilde{D}_\theta(u)$.

Another consequence of considering information of order α is that we can unify the estimation criterion used to form maximum likelihood estimators with the estimation criterion used to form Gaussian time series parameter estimators:

$$I_{sp}(\tilde{f};f_\theta) = \log \int_0^1 \frac{\tilde{f}(w)}{f_\theta(w)}\, dw ,$$

where $\tilde{f}$ and f_θ are spectral densities. It is comparable to

$$I_2(\tilde{d}_\theta) = \log \int_0^1 \frac{\tilde{f}\tilde{Q}(u)}{f_\theta \tilde{Q}(u)}\, du$$

7. Statistical inference reduced to density estimation

The quantile approach to statistical data analysis being developed by Parzen [since Parzen (1979)] is based on the proposition that conventional problems of statistical inference concerning (1) a random sample $X_1,\ldots,X_n$, (2) a bivariate sample $(X_1,Y_1),\ldots,(X_n,Y_n)$, or (3) two samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ should be transformed to problems of functional inference, estimating and testing hypotheses about density functions $d(u)$, $d(u_1,u_2),\ldots,d(u_1,\ldots,u_k)$, on the unit interval $0\leq u\leq 1$, unit square $0\leq u_1,u_2\leq 1$, unit hypercube $0\leq u_1,\ldots,u_k\leq 1$. To illustrate how this is done consider the following problems.

Modeling Bivariate Data and Tests for Indpenedence. Let X and Y be continuous random variables with joint density function $f_{X,Y}(x,y)$. The hypothesis, Ho: X and Y are independent can be expressed

$$\text{Ho: } f_{X,Y}(x,y) = f_X(x)\ f_Y(y)$$

or in terms of information divergence

$$I(f_{X,Y};f_Xf_Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{-\log \frac{f_X(x)f_Y(y)}{f_{X,Y}(x,y)}\}\ f_{X,Y}(x,y)\ \ dx\ dy$$

by

$$\text{Ho: } I(f_{X,Y};\ f_Xf_Y) = 0\ .$$

Define

$$D(u_1,u_2) = F_{X,Y}(Q_X(u_1),Q_Y(u_2))$$

$$d(u_1,u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} D(u_1,u_2)$$

$$= \frac{f_{X,Y}(Q_X(u_1),Q_Y(u_2))}{f_X Q_X(u_1)\ f_Y Q_Y(u_2)}$$

We call $d(u_1,u_2)$ the quantile dependence density.

The hypothesis Ho can be expressed

$$\text{Ho: } D(u_1,u_2) = u_1u_2,\ d(u_1,u_2) = 1.$$

One can verify that

$$I_1(f_{X,Y};f_Xf_Y) = \int_0^1 \int_0^1 \{\log d(u_1,u_2)\}\, d(u_1,u_2)\, du_1du_2$$

$$= - H_1(d(u_1,u_2))$$

Thus estimating the information divergence between $f_{X,Y}$ and f_Xf_Y is equivalent to estimating the negative of the entropy of $d(u_1,u_2)$.

Estimators $\hat{d}_m(u)$ dependent on a finite number of parameters can be formed from the raw estimator

$$\tilde{D}(u_1,u_2) = \tilde{F}_{X,Y}(\tilde{Q}_X(u_1),\ \tilde{Q}_Y(u_2)).$$

Modeling likelihood ratios and testing equality of distributions. Let X and Y be continuous random variables. The hypothesis

$$\text{Ho: } F_X(x) = F_Y(x), \text{ or } f_X(x) = f_Y(x)$$

can be expressed in terms of information divergence

$$I(f_Y;f_X) = \int_{-\infty}^{\infty} -\log \frac{f_X(x)}{f_Y(x)} \, dF_Y(y)$$

$$= \int_0^1 -\log d(u) \, du$$

$$= -Hqd\,(d(u))$$

defining the comparison distribution function and comparison density function

$$D(u) = F_X Q_Y(u), \quad d(u) = \frac{d}{du} D(u) = \frac{f_X(Q_Y(u))}{f_Y(Q_Y(u))}$$

Estimating the information divergence between f_Y and f_X is equivalent to estimating the negative of the entropy in the quantile-density sense of the comparison density d(u).

8. Parametric-select density estimation and Maximum Entropy Densities

A density $d(u) = D'(u)$ can be approximated in many ways by sequences $d_m(u), m=1,2,\ldots$ of functions which converge to d(u). For $m=1,2,\ldots$, let $\hat{d}_m(u)$ be an estimator of $d_m(u)$; the sequence $\hat{d}_m(u)$ then estimates d(u).

If $d_m(u)$ corresponds to a standard finite parameteric model d(u) for which one could consider testing the hypothesis that $d_m(u)$ provides an exact model, we call $d_m(u)$ a parametric-select representation, and $\hat{d}_m(u)$ a parametric-select estimator,

to indicate that we are free to select the number of parameters in $d_m(u)$ to provide an adequate approximation or representation of $d(u)$.

We call $d_m(u)$ a non-parametric representation, and $\hat{d}_m(u)$ a non-parametric estimator, if $d_m(u)$ does not correspond to a standard finite parameter model which could be interpreted as an exact model.

An important criterion for developing the functional form of exact models for densities is the maximum entropy principles.

A density $f(x)$, $-\infty<x<\infty$, which maximizes entropy $H(f) = \int_{-\infty}^{\infty}\{-\log f(x)\}f(x)\,dx$ subject to constraints

$$\int_{-\infty}^{\infty} T_j(x)\, f(x)\, dx = \tau_j, \quad j=1,\ldots,k,$$

where $T_j(x)$ are specified functions (called sufficient statistics) and τ_j are specified moments can be shown to have the representation, called an exponential model,

$$\log f(x) = \sum_{j=1}^{k} \theta_j\, T_j(x) - \psi(\theta_1,\ldots,\theta_k)$$

where

$$\psi(\theta_1,\ldots,\theta_k) = \log \int_{-\infty}^{\infty} \exp\{\sum_{j=1}^{k} \theta_j\, T_j(x)\}\, dx$$

guarantees that $f(x)$ integrates to 1.

A quantile function $q(u)$, $0<u<1$, which maximizes entropy $Hqd(q) = \int_0^1 \log q(u)\, du$ subject to the constraints

$$\frac{\int_0^1 \exp\,(2\pi iuv)\; f_oQ_o(u)\; q(u)\; du}{\int_0^1 f_oQ_o(u)\; q(u)\;\; du} = \rho(v),\; v=0,\pm1,\ldots,\pm m$$

where $f_oQ_o(u)$ is a specified density quantile function must have the representation, called an autoregressive model,

$$q(u) = q_o(u)\; \sigma_m^2\; |1+\alpha_m(1)e^{2\pi iu}+\ldots+\alpha_m(m)e^{2\pi ium}|^{-2}$$

9. Exact-Parametric and Parameter-select Estimation of Probability density Functions using Exponential Models

Two important exponential models for a density f(x), $-\infty<x<\infty$ are the normal density and the gamma density.

The normal density, denoted Normal (μ,σ)

$$f_{\mu,\sigma}\;(x) = \frac{1}{\sigma}\,\phi(\frac{x-\mu}{\sigma})\;,$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp -\frac{1}{2}\,x^2$$

is exponential with sufficient statistics $T_1(x) = x$ and $T_2(x) = x^2$.

The Gamma density, denoted Gamma (r,λ) where $\lambda = 1/\sigma$,

$$f_{r,\sigma}\;(x) = \frac{1}{\sigma}\,f_r\;(\frac{x}{\sigma})\;,$$

$$f_r(x) = \frac{1}{\Gamma(r)}\,x^{r-1}\,e^{-x}\quad,\quad x>0\;,$$

$$= 0\qquad\qquad\quad,\quad x<0\;,$$

is exponential with sufficient statistics $T_1(x) = x$ and $T_2(x) = \log x$.

A location scale parameter Gamma density

$$f_{r,\mu,\sigma}(x) = \frac{1}{\sigma} f_r(\frac{x-\mu}{\sigma})$$

is not an exponential model. We can treat it as one by estimating μ (say, by the minimum $X_{(1)}$ of the random sample $X_1,\ldots,X_n$), and treating $X_j-\hat{\mu}$ as a sample from $f_{r,\sigma}(x)$.

The hypothesis that the data is fit by a normal distribution versus the hypothesis that the data is fit by a Gamma distribution can be tested by forming an over-parametrized exponential model with sufficient statistics

$$T_1(x) = x, \quad T_2(x) = x^2, \quad T_3(x) = x^3, \quad T_4(x) = \log x.$$

The (order 1) information divergence, or maximum likelihood, estimators $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, $\hat{\theta}_4$, which minimize information divergence of order 1 $\int_0^1 -\log \tilde{d}_\theta(u)\, du$, may be found for an exponential model by solving

$$\hat{\tau}_j = E_{\hat{\theta}}[T_j]$$

where $\tau_j = E_\theta[T_j]$ is estimated by

$$\hat{\tau}_j = \bar{T}_j = \frac{1}{n} \sum_{j=1}^{n} T_j(X_{(j)}) \quad .$$

The bi-information divergence estimators $\check{\theta}_1$, $\check{\theta}_2$, $\check{\theta}_3$, $\check{\theta}_4$, which minimize information divergence $\int_0^1 |\log \tilde{d}_\theta(u)|^2 du$, may be found using least squares regression analysis techniques by minimizing with respect to $\theta_1,\ldots,\theta_k$ the sum of squares

$$\sum_{j=2}^{n-1} |\log \tilde{f}(X_{(j)}) - \{\log \tilde{f}(X_j)\}^- - \theta_1 (T_1(X_{(j)}) - \bar{T}_1) - \ldots - \theta_k (T_k(X_{(j)}) - \bar{T}_k)|^2$$

Stepwise regression is used to suggest parsimonious parametrizations.

Graphical procedures to determine which parameter values fit best are as follows: estimate $\tilde{D}_\theta(\frac{j}{n+1})$, $j=2,\ldots,n-1$, by adding

$$\tilde{d}_\theta(\tfrac{j}{n+1}) = f_\theta(X_{(j)}) \div \tilde{f}(X_{(j)})$$

and normalizing the sum to go from 0 to 1. One inspects its graph to see how it deviates from $D(u) = u$.

10. Case studies of bi-information density estimation

The density estimators corresponding to the bi-information parameter estimates of the normal, gamma, and four-parameter exponential models are presented for four simulated random samples:

1) Exponential or Gamma ($r = 1$, $\sigma = 1$)

2) Gamma ($r=10$, $\sigma =1$)

3) Normal ($\mu = 0$, $\sigma = 1$),

4) Contaminated normal: 100N(0,1),5N(10,1)

In addition density estimators, using bi-information parameters, are presented for the data set of Buffalo snowfall. Bi-information select regression estimation of the parameters of a 4-paramential exponential model with sufficient statistics x, x^2, x^3, and log x leads to the conclusion that Buffalo snowfall obeys a Gamma distribution. It is equally well fit by a normal distribution whose parameters are estimated by minimizing bi-information rather than order 1 information. The hypothesis that Buffalo snowfall is normal seems to be acceptable, but one can question whether the maximum likelihood estimators (sample mean and variance) provide the best-fitting normal distribution for Buffalo snowfall.

As in Parzen (1979), we reject a trimodal shape probability density estimate for Buffalo snowfall, which has been found by several non-parametric density estimation techniques; including Tapia and Thompson (1978).

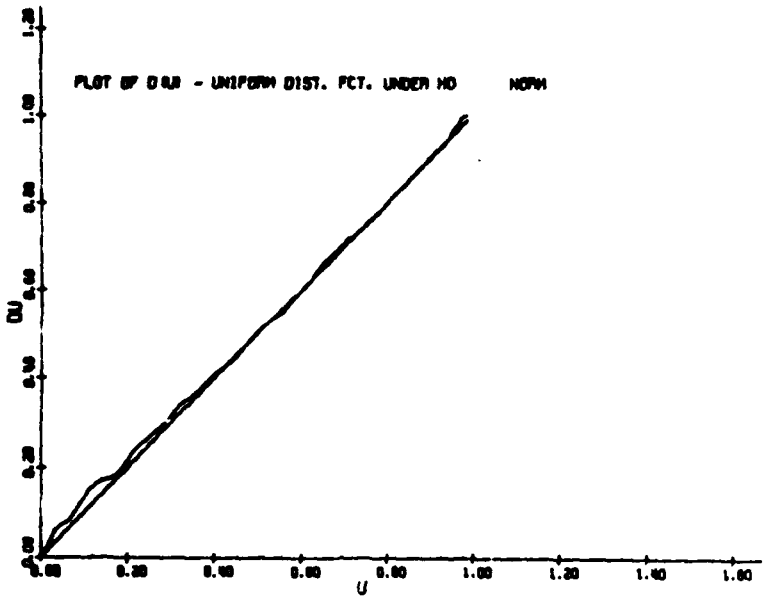
REFERENCES

Parzen, E. (1979) Nonparametric statistical data modeling. Journal of the American Statistical Association, 74, 105-131.

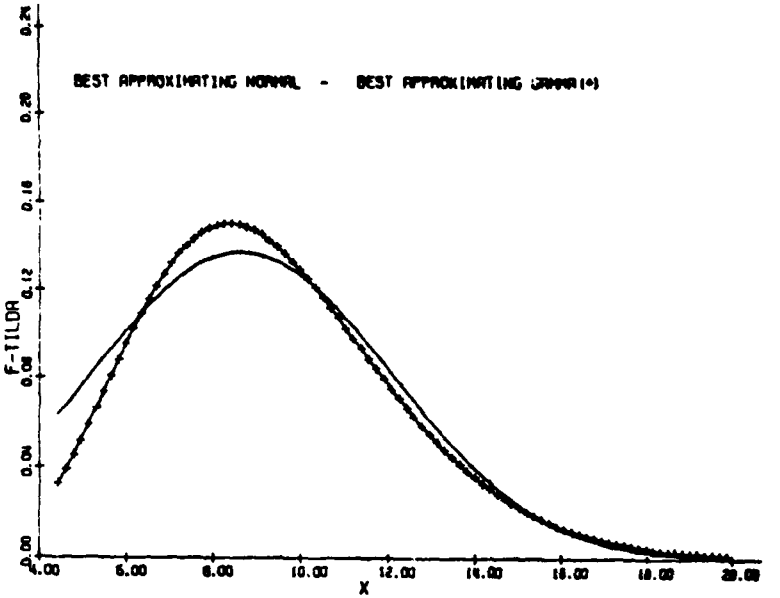
_____. (1982) Maximum entropy interpretation of autoregressive spectral densities. Submitted for publication.

Tapia, R. A. and Thompson, J. R. (1978) Nonparametric Probability Density Estimation, Baltimore: Johns Hopkins University Press.

$\tilde{D}(u)$ FOR BI-INF-DIV NORMAL (8.64, 11.53)

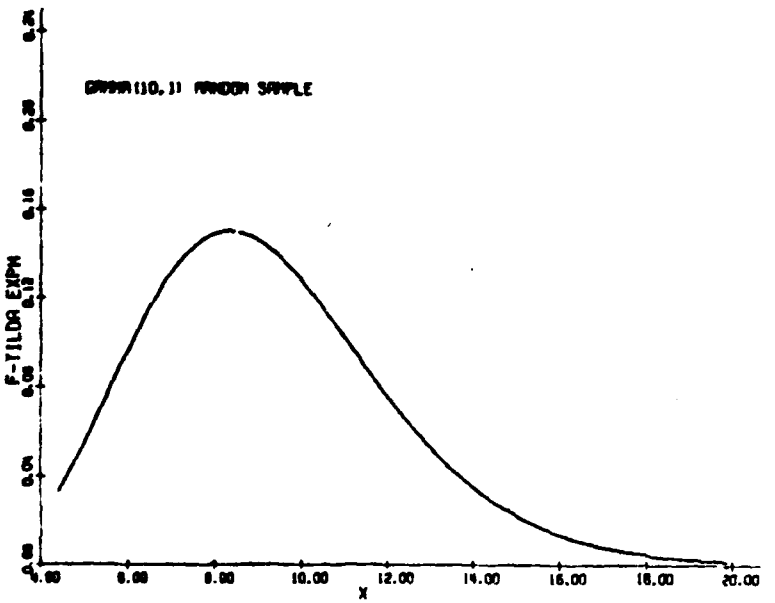


DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE

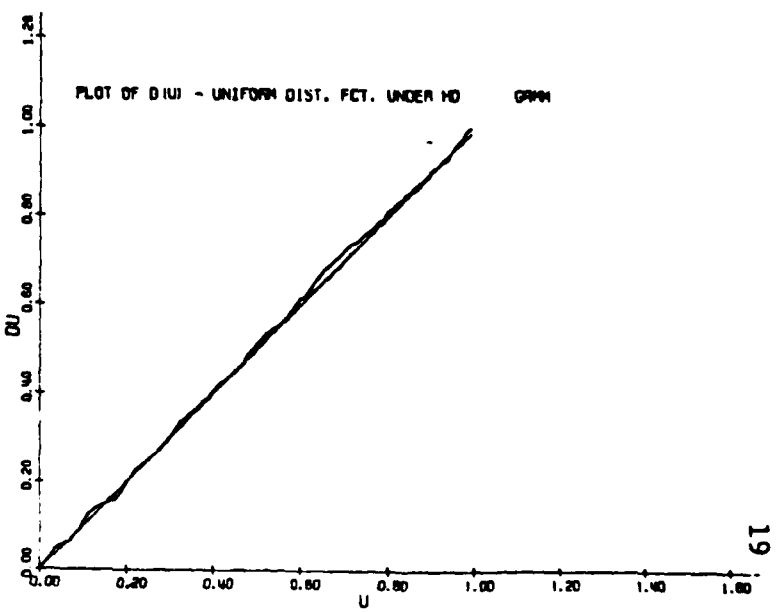


GAMMA (10,1) SIMULATED SAMPLE

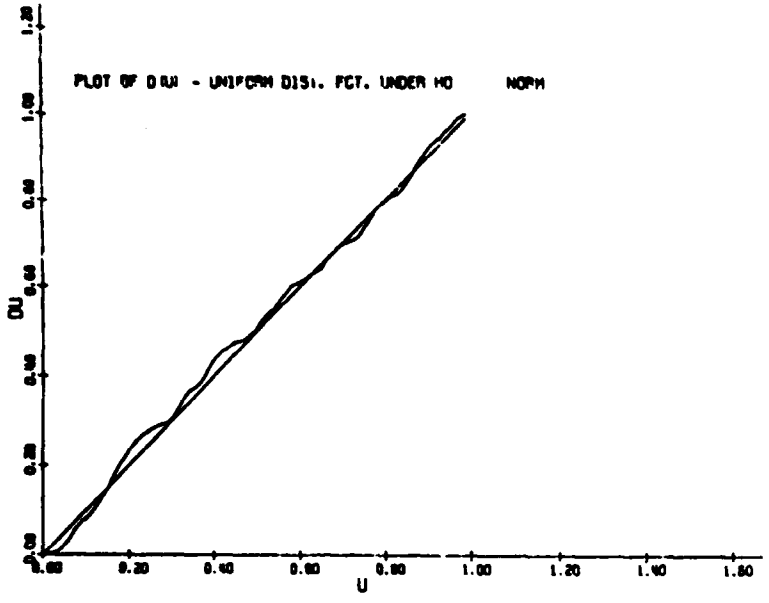
PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS GAMMA (10.08, 1.09)



$\tilde{D}(u)$ FOR BI-INF-DIV GAMMA (10.08, 1.09)

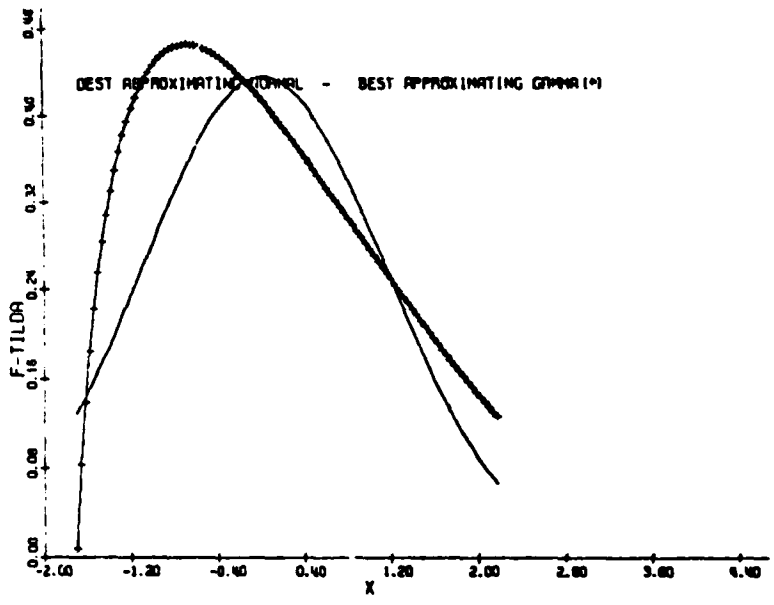


$\tilde{D}(u)$ FOR BI-INF-DIV NORMAL (.03, 1.22)

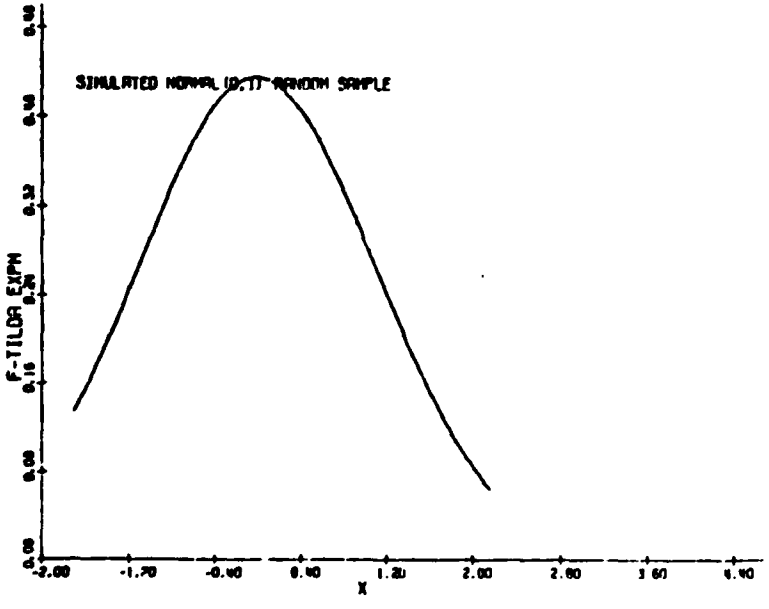


NORMAL (0, 1) SIMULATED SAMPLE
Sample Mean .11, Variance .82

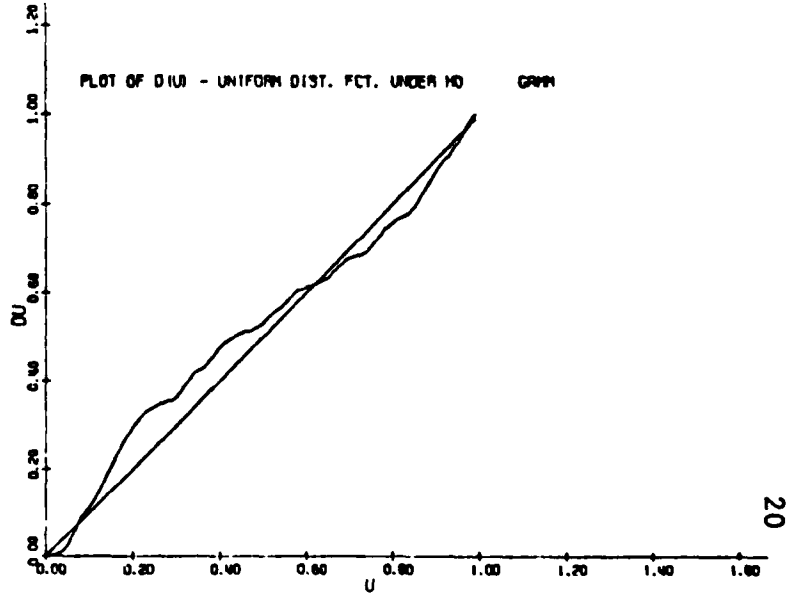
DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE



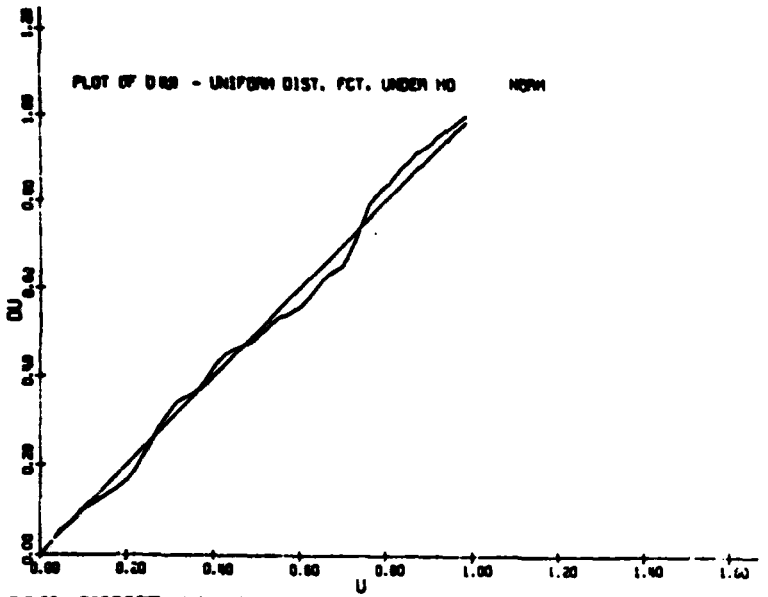
PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS NORMAL (.03, 1.22)



$\tilde{D}(u)$ FOR BI-INF-DIV GAMMA (1.56, .53)

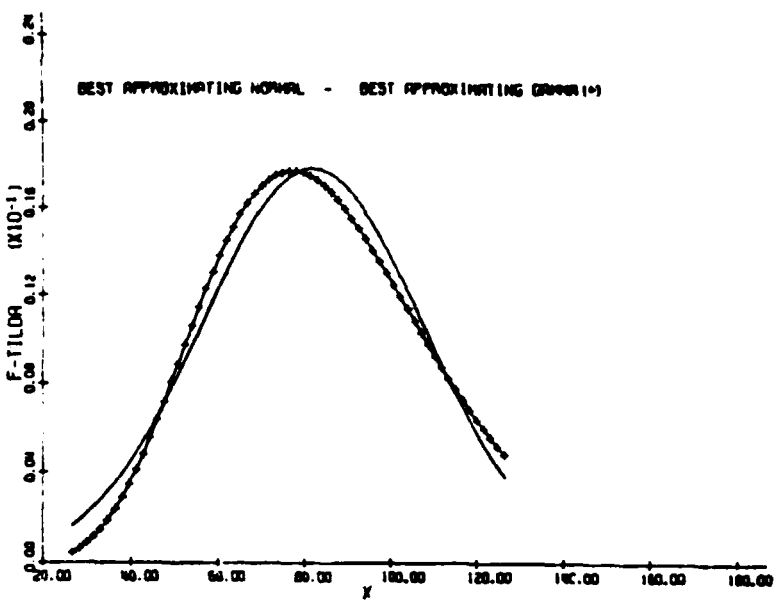


$\tilde{D}(u)$ FOR BI-INF-DIV NORMAL (81.9, 644)

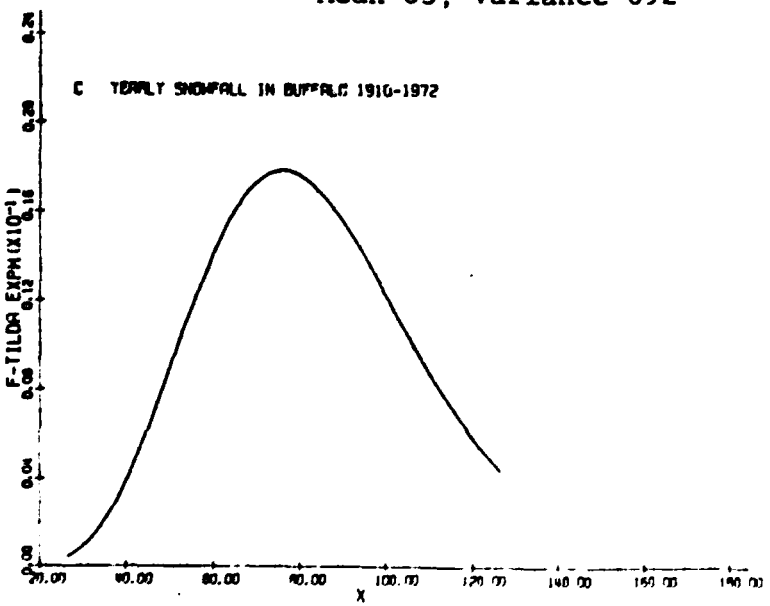


YEARLY SNOWFALL IN BUFFALO 1910-1972
Mean 80.5, Variance 487

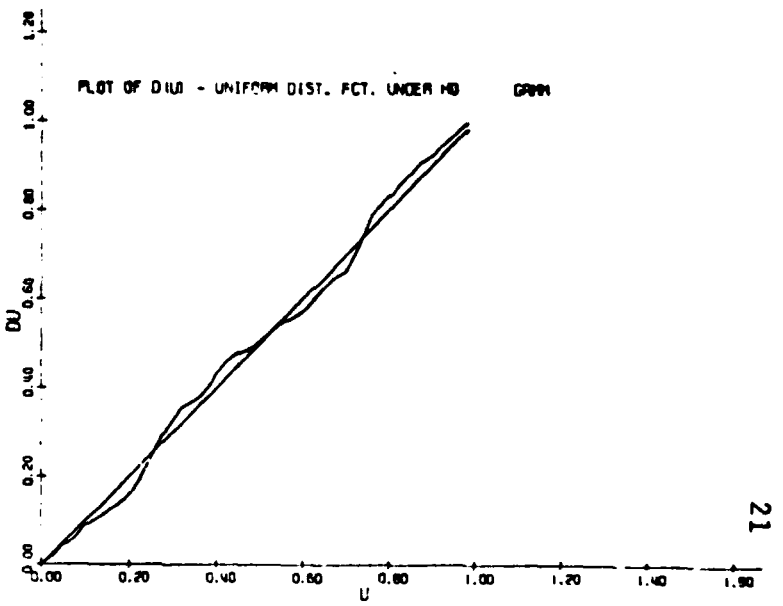
DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE



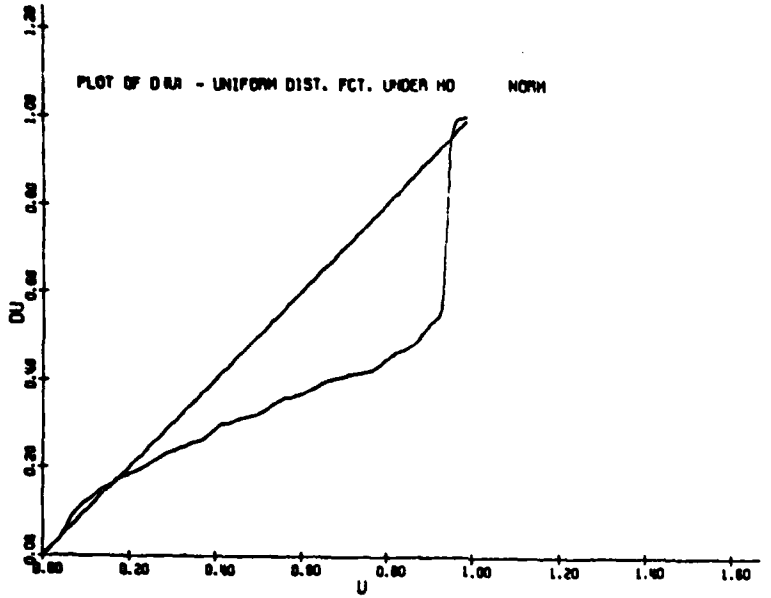
PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS GAMMA (9.96, .12) Mean 83, Variance 692



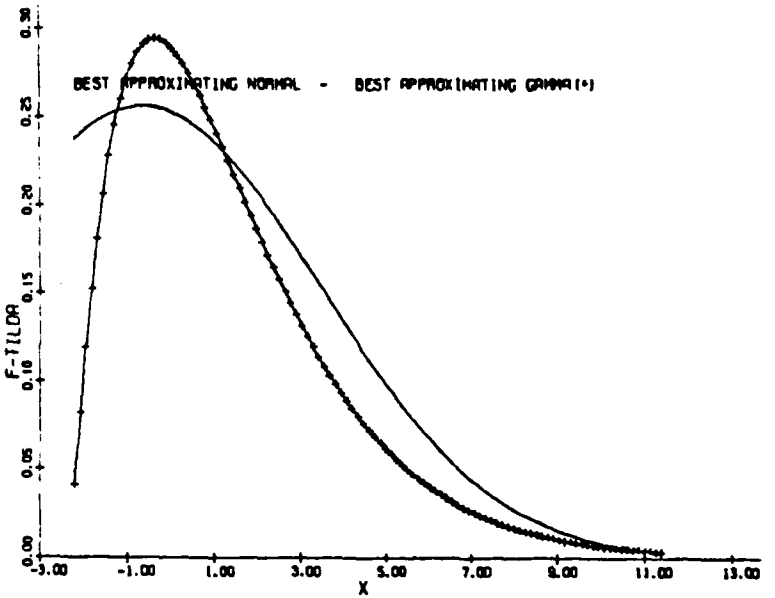
$\tilde{D}(u)$ FOR BI-INF-DIV GAMMA (9.96, .12)



D(u) FOR BI-INF-DIV NORMAL (-.63, 16.51)

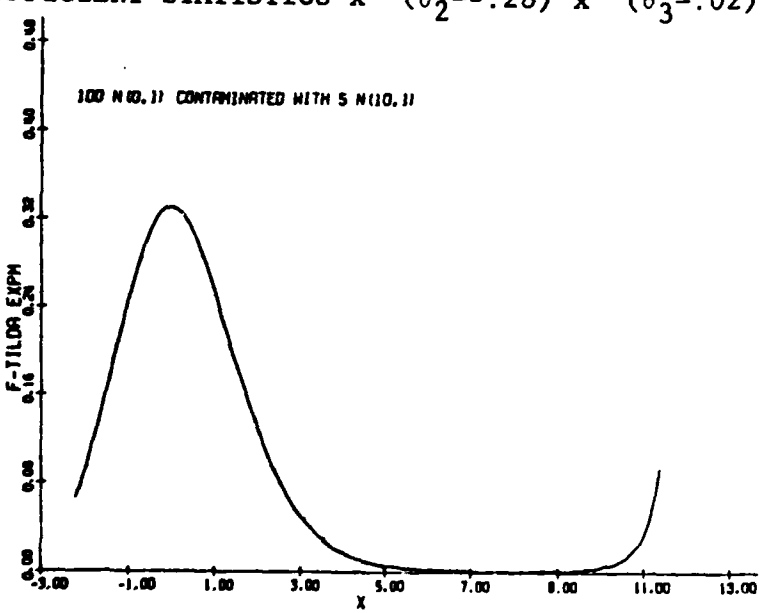


DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE

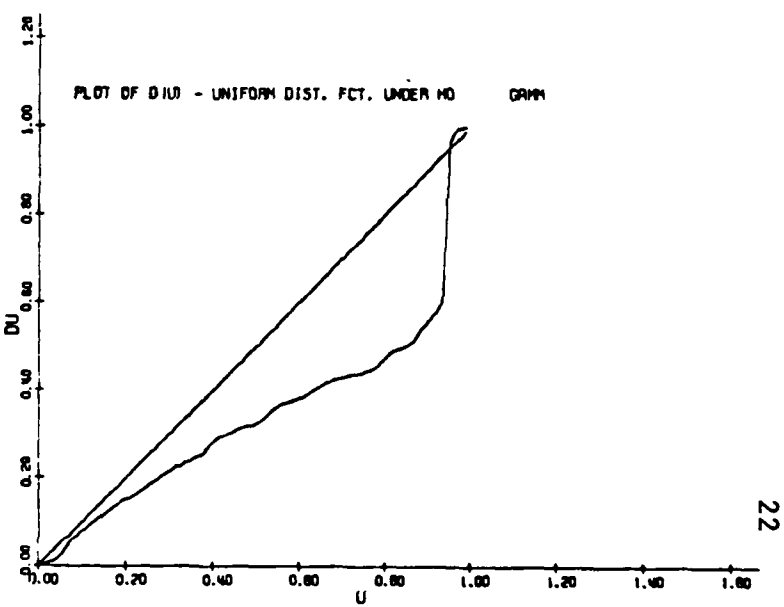


CONTAMINATED NORMAL 100N(0,1), 5N(10,1)
Mean .4, Variance 4.6.

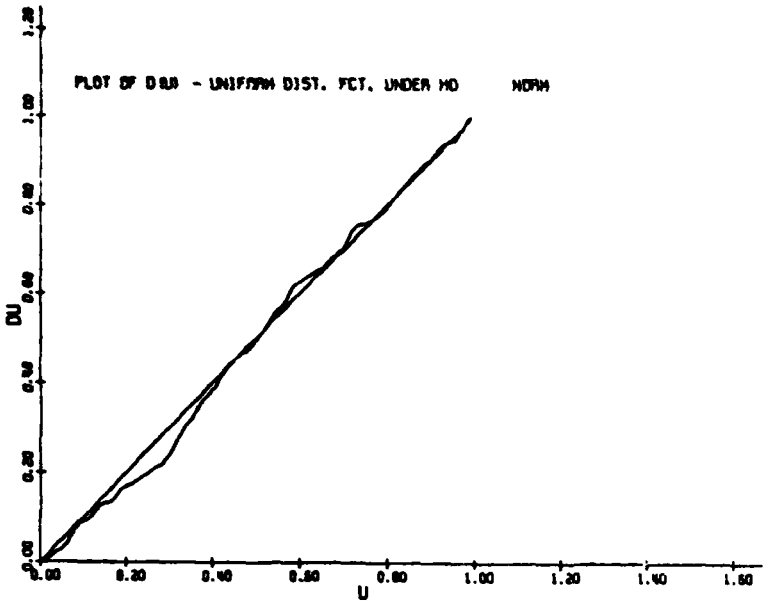
PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY HAS SUFFICIENT STATISTICS x^2 (θ_2=-.28) x^3 (θ_3=.02)



D(u) FOR BI-INF-DIV GAMMA (2.11, .57)

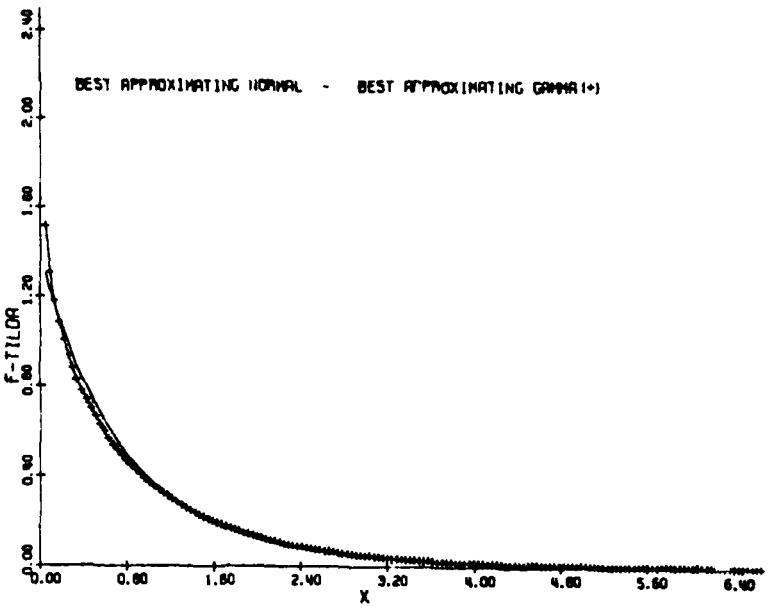


$\tilde{D}(u)$ FOR BI-INF-DIV NORMAL (11.4,-8.6)

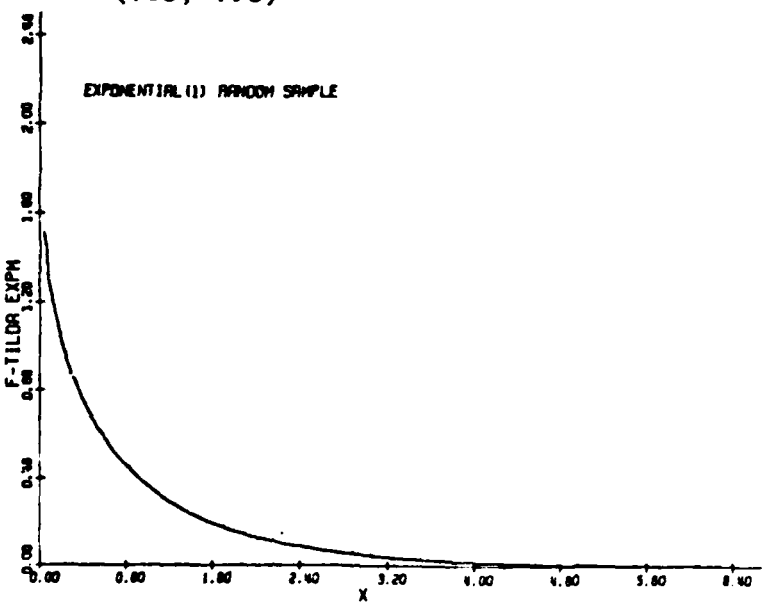


EXPONENTIAL GAMMA (r=1, λ=1) SIMULATED SAMPLE

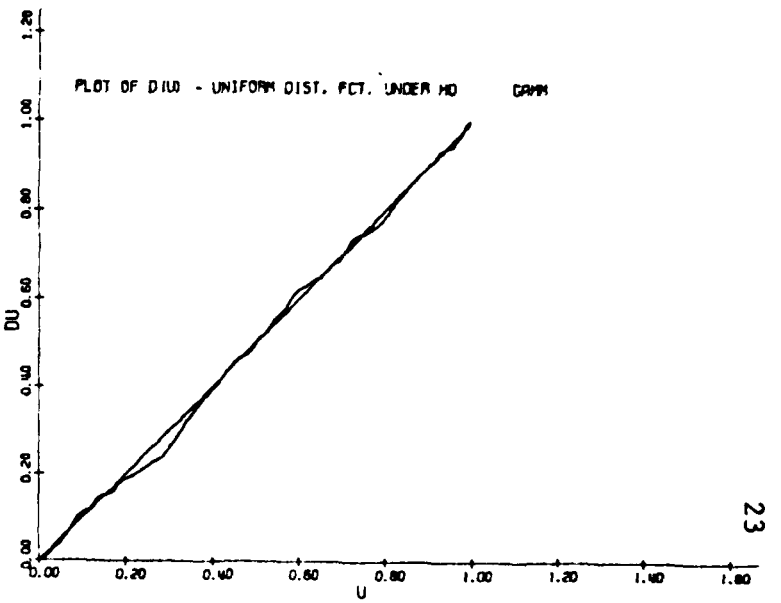
DENSITY ESTIMATES EVALUATED ON SAMPLE RANGE



PARSIMONIOUS 4-PARAMETER EXPONENTIAL DENSITY IS GAMMA (.83, .93)



$\tilde{D}(u)$ FOR BI-INF-DIV GAMMA (.83, .93)




ATE
MED